

Improved half-maximal inhibitory concentration regression model using amyotrophic lateral sclerosis data

Devipriya Selvaraj¹, Vijaya M S², Krishnaveni Sakkarapani³

¹Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India

²Department of Computer Science (PG), PSGR Krishnammal College for Women, Coimbatore, India

³Department of Data Analytics (PG), PSGR Krishnammal College for Women, Coimbatore, India

Article Info

Article history:

Received Mar 29, 2024

Revised Feb 13, 2025

Accepted Mar 9, 2025

Keywords:

Amyotrophic lateral sclerosis

Deep learning

Drug discovery

Gene expression

Regression

ABSTRACT

The current research addresses the critical need for precise half-maximal inhibitory concentration regression in the neurodegenerative condition amyotrophic lateral sclerosis (ALS). Unavailable drug-induced gene expressions and irrelevant molecular descriptors have yielded regression models with less accuracy using traditional machine learning (ML). Drugs can be converted to graph format and integrated with gene expressions to learn drug-gene interactions better thereby producing precise half-maximal inhibitory concentration regression models. To accomplish this, three variants of graph neural networks (GNN) namely graph attention networks (GAT), message passing neural networks, and graph isomorphism networks are utilized in the proposed work. The gene expression profiles of ALS drug-related genes were retrieved from the DepMap PRISM drug repurposing hub, and the drug graphs with their accompanying half-maximal inhibitory concentration values were obtained from the ChEMBL databases. The graph is constructed for ninety approved drugs connected to 32 key protein targets of ALS and its related conditions. The half-maximal inhibitory concentration regression model trained with optimized hyperparameters in GAT performs well with an R2 score of 0.92, a mean absolute error (MAE) of 0.20, and a root mean square error (RMSE) of 0.17. This model produced better results than other ML and deep learning models.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Devipriya Selvaraj

Department of Computer Science, PSGR Krishnammal College for Women

Coimbatore, India

Email: devipriya041996@gmail.com

1. INTRODUCTION

The process of predicting the problem of half-maximal inhibitory concentration (IC50) in amyotrophic lateral sclerosis (ALS) is a complex task including exhaustive methods of identifying, and improving drugs that can be used in disease treatment [1]. ALS [2], [3] is a degenerative neurological disorder that affects the nerve cells responsible for voluntary muscle movement. Several important targets require inhibition in ALS due to the mutations involved in them, like superoxide dismutase 1 (SOD1) [4], TAR DNA-binding protein 43 (TDP-43) [5], C9orf72 repeat expansion [6], fused in sarcoma (FUS) [7], proteostasis and protein quality control [8], glutamate excitotoxicity [9], neuroinflammation, mitochondrial dysfunction [10], axonal transport [11], RNA metabolism and processing [12]. Precise prediction of IC50 values is essential for understanding and combating the biochemical mechanisms underlying ALS and directly aiding in drug discovery [13]-[16]. The analysis of machine learning (ML) and deep learning algorithms is thoroughly discussed in this research.

Early quantitative structure-activity relationship (QSAR) [17] technologies for IC₅₀ prediction lacked versatility and relied heavily on traditional ML methods using 2D and 3D molecular descriptors and expert interpretation. Integrating big data with advanced ML algorithms has significantly enhanced QSAR models' ability to handle unstructured data, thereby improving the process of discovering new drugs. The review highlights the evolution of QSAR techniques, combining wet lab experiments, molecular dynamics simulations, and ML approaches. The importance of merging data inputs to optimize drug development is emphasized. A study [18] outlines an automated framework for QSAR model building, automating tasks like data curation, feature selection, and validation, leading to 19% reduced error and 49% increased explained variance. Another paper [19] explores meta-learning, improving QSAR prediction accuracy by up to 13% across 2,700 datasets, showing that random forests with fingerprints as an alternative for molecular descriptors often perform best. The final [20] study finds no universally superior algorithm for QSAR but notes that non-linear methods, such as radial basis function support vector machine (RBF SVM), extreme gradient boosting (XGBoost), and deep learning algorithms, generally outperform linear ones, with ensemble methods providing further enhancements.

The regression of IC₅₀ values relies on access to high-quality, annotated datasets. The data includes both molecular and biological information. Integrating extensive datasets that include various targets, compounds, their interactions, and disease data, such as gene expression, into ML algorithms is challenging. Therefore, it has been overcome and prompted research into the utilization of deep learning in ALS IC₅₀ problems.

The paper by Deng *et al.* [21] has made a significant impact in the area of neurodegenerative diseases. The dataset utilized in the research comprised substances sourced from Chinese medicine databases and the ZINC database. The authors employed various artificial intelligence techniques, specifically deep learning methods and XGBoost models, to identify molecules that bind to the target protein galectin 3. Galectin 3 has been used as a protein that requires inhibition. The numerous modeling attempts by the deep learning-based algorithm and the obtained R-square correlation coefficient of 0.9 on test sets demonstrated its effectiveness. The XGBoost model produces a 0.97 R-square correlation coefficient and 0.01 mean square error. The study by Knutson *et al.* [22] utilized the novel parallel graph neural networks (GNN) technique to predict and assess complex chemical interactions. The thorough analysis showed that GNN can accurately predict the protein-ligand complex activity by capturing the binary interactions that existed between them, with test accuracy for GNN for ligand-feature interactions (GNNF) being 0.979 and for GNN for protein-feature interactions (GNNP) being 0.958. The interactions between proteins and ligands have been enhanced through GNNs and parallel computing.

Sagingalieva *et al.* [23] developed a distinctive computational approach for predicting drug responses by integrating traditional ML with quantum information processing elements. The hybrid quantum neural network (HQNN) model evolved as a fresh and possibly revolutionary technique in drug response prediction, leveraging the computational capacity of quantum bits (qubits) for specialized calculations. The study utilized the genomics of drug sensitivity in cancer (GDSC) dataset, which offers data on the responses of cancer cell lines to anticancer drugs, along with genetic associations. The study discovered that the HQNN outperformed standard models, with a 15% lower error rate in predicting IC₅₀ values. Zuo *et al.* [24] described a strategy that combines compound chemical structures with cancer genetic signatures to predict drug response. The study used the SWnet technology, which employs convolutional and recurrent neural network architectures for deployment. The model provides a flexible and adaptive approach to detecting drug interactions. The incorporation of attention techniques and feature fusion increased architectural performance. The parameters provided were accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) paired with and without attention mechanisms. The architecture using attention mechanisms yielded higher precision.

One of the key contributions to personalized medicine was made by Park *et al.* [25]. The study compared various models that predicted the IC₅₀ values for cell viability of 24 distinct drugs, utilizing gene expression and mutation data from cancer cell lines. The EC-11K gene expression and the MC-9K mutation dataset were trained for building drug response prediction models. Deep learning models utilized convolutional neural networks (CNN), ResNet architectures, and ML models such as lasso, ridge, support vector regression (SVR), random forest, XGBoost, and ElasticNet. The researchers assessed the models' performance using visual inspection, R², and root mean squared. Among the 24 personalized drugs, the ridge model of Panobinostat proved to be the best with an R² score of 0.470 and RMSE of 0.623.

In the previous work, ML-based IC₅₀ prediction models were built with inappropriate drug descriptors. Many GNNs algorithms for building IC₅₀ prediction models employ drug-induced gene expression which produces less prediction accuracy. Thus, the proposed methodology uses drug simplified molecular input line entry system (SMILES) and gene expression as global features with various advanced GNN architectures graph attention network (GAT), message passing neural network (MPNN), and graph isomorphism network (GIN) for building IC₅₀ regression models. The proposed work combines drug

Improved half-maximal inhibitory concentration regression model using amyotrophic ... (Devipriya Selvaraj)

SMILES and gene expressions with advanced deep learning algorithms to construct an accurate IC50 prediction model that effectively learns drug-gene interactions.

This research paper is further arranged as follows: the method is presented in section 2 with various processes namely data collection, data transformation, and IC50 regression model building process. At the same time, the results and discussion of the model are demonstrated in section 3. At last, the conclusion of this research is given in section 4.

2. METHOD

The method uses the novel methodology of integrating drug SMILES with gene expressions and passing them to GNN variants. Figure 1 represents the flow of the research. It starts with data collection of ninety drug graphs from ChEMBL and gene expressions from DepMap PRISM drug repurposing hub for ninety drug-related genes. The collected data is transformed into feature and adjacency matrices, followed by k-fold cross-validation with a train-test data split. Three GNN variants namely GAT [26], MPNN [27], and GIN [28] are used for the IC50 regression model building process. The model is evaluated with three metrics mean absolute error (MAE), root mean square error (RMSE), and R2 score.

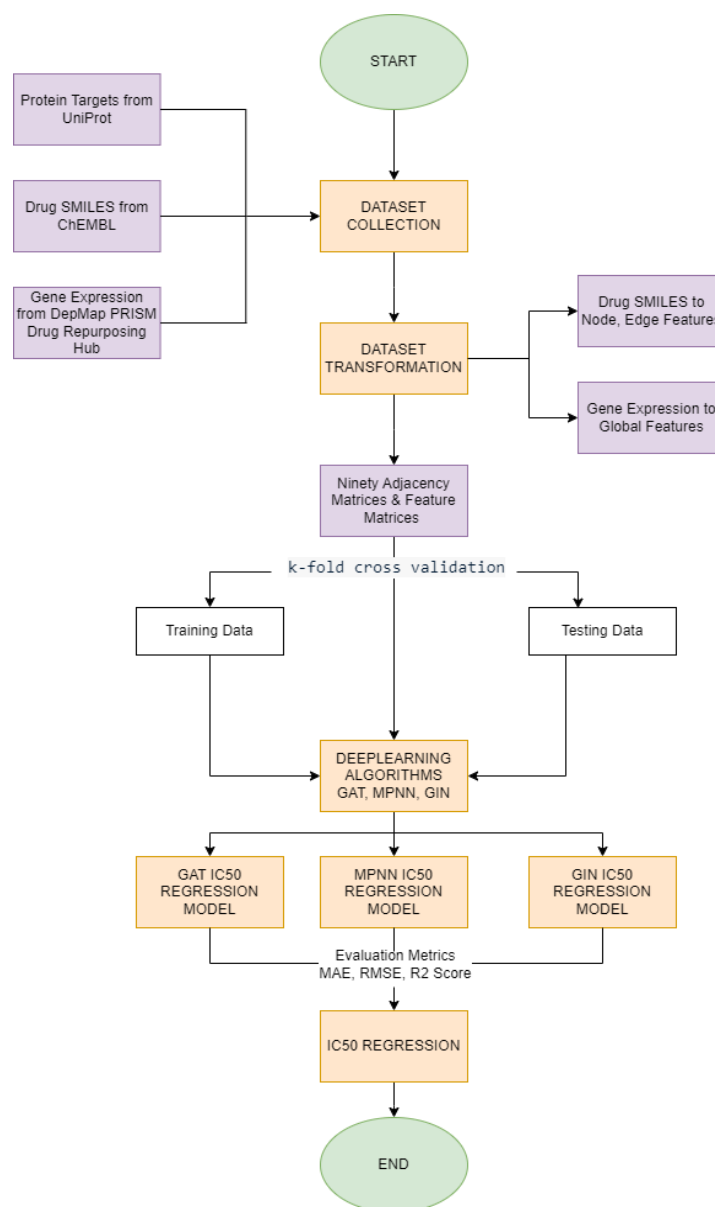


Figure 1. Flow of research

2.1. Dataset collection and transformation

The data is collected from different sources like UniProt [29], ChEMBL [30], and DepMap [31]. The UniProt database is searched for the ALS targets, and they are retrieved. It is discovered through pathway analysis that all 32 protein targets are present in ALS. The ChEMBL database is used to identify the drugs linked to the 32 ALS targets and the associated SMILES of drugs. Only ninety authorized drugs are selected from the list of compounds compiled through ChEMBL target mapping.

Drugs and their targets associated with ALS are identified using pathway analysis [29], [32] as shown in Figure 2. Pathway analysis is employed to clarify complex biological processes such as ALS by identifying interconnected molecular pathways. In the figure, only the best results are deployed where the adjusted P-value is lower and the Effect size is larger. The drug SMILES and pIC50 values are obtained from the ChEMBL database through target mapping to find the drug characteristics. Drug SMILES offer a condensed and standardized method of employing a certain collection of characters to indicate the structure of a drug molecule. IC50 quantifies the potency or inhibitory effect of a medicinal molecule. pIC50 is the negative logarithm of base 10 of the medication concentration needed to 50% block a target. The pIC50 is a normalized value of IC50 frequently utilized in drug discovery. It ranges from 0 to 10 where values near 10 indicate better drugs. The drugs having values above or below the range are not used for model building.

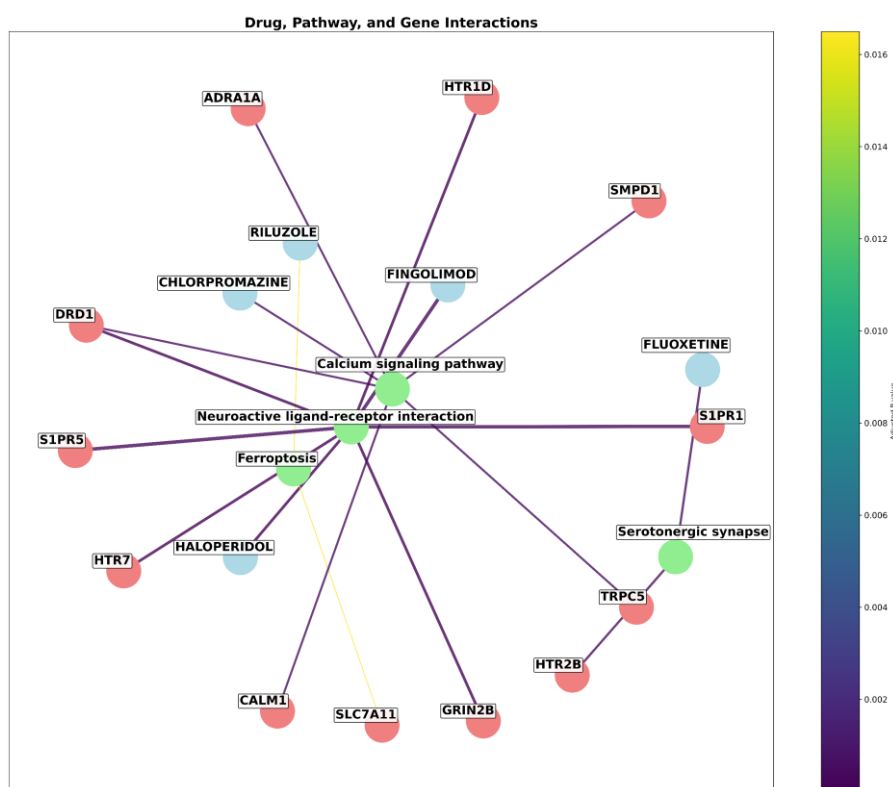


Figure 2. Pathway analysis

The DepMap database has cellular gene expression data from which 90 drugs studied about ALS are collected. DepMap is a large database that provides transcriptome, the genome, and variations in the genome for disease-related cell lines. A sample gene expression is shown in Table 1. Genes with positive values signify that these genes are inhibited by drugs and are expressed at levels higher than the baseline. The higher the positive value, the higher the expression level relative to the baseline. Genes with negative values signify that growth is induced by drugs and are expressed at levels lower than the baseline are removed. The lower the negative value, the lower the expression level relative to the baseline. A value of 0 indicates that the gene is expressed at the baseline level. The DepMap database receives the query and identifies ALS-related cell lines in the PRISM drug repurposing center. The PRISM drug repurposing hub in DepMap is the basis for the drug repurposing. As a result, information regarding the expression levels of particular genes in ALS-related cell lines is retrieved. The values signify the need for inhibitory characteristics of drug targets in the ALS disease.

Table 1. Sample gene expression

depmap_id	cell_line_display_name	lineage_1	Expression public 23Q2 KCNA1	Expression public 23Q2 KCNA10	Expression public 23Q2 KCNA3	Expression public 23Q2 KCNA2	Expression public 23Q2 KCNA4
ACH-000102	GMS10	CNS/Brain	0	0.028569	0.028569	0	0.014355
ACH-000200	NMCG1	CNS/Brain	0	0	0.014355	0	0
ACH-000504	SNB75	CNS/Brain	0	0.124328	0	1.144046	0
ACH-001211	TTC549	CNS/Brain	0	0	0	0.070389	0
ACH-000597	TTC709	CNS/Brain	0	0	0	0	1.111031
ACH-000623	SNU201	CNS/Brain	0	0	0	0.189034	0.028569
ACH-000543	SNU489	CNS/Brain	0.137504	0	0	0.056584	0

During the dataset transformation, a featurization procedure is used, drug SMILES are converted into feature vectors and the gene expressions are standardized. The three different types of feature vectors are node, edge, and global feature vectors. The node features include attributes like atom type, formal charge, hybridization, and aromaticity. The edge features have attributes of bond type, bond distance, and graph distance. Gene expression and drug SMILES work together to create feature vectors. The gene expression for each drug is standardized using Min-Max normalization and added to the corresponding drug SMILES graph feature vector. These gene expression data are added to the graph feature of each drug in addition to the node and edge features. Gene expression in different cell lines is utilized as a feature to determine the impact of the drug on them. Gene expression data relating to each drug SMILES is provided as a global feature vector. The adjacency matrix is provided as input to the model to find graph structure. The purpose of the adjacency matrix is to provide details on the structure of drug molecules such as the presence of atoms, bonds, and global features. Global features can be incorporated by adding additional rows and columns to the adjacency matrix.

Following featurization, the feature extraction procedure happens in various GNN topologies. This process varies in GNN variants based on its architecture producing node embeddings, where information is iteratively exchanged between neighboring atoms in iterations. This information exchange enables the refinement of node representations by incorporating insights from neighboring atoms. As a result of feature extraction, the feature representation of each atom becomes enriched with contextual information about its surrounding atoms and bonds. Aggregate node-level features are converted into a fixed-size vector, thereby capturing the collective characteristics of the entire molecule. It is usually done by a readout function producing molecule-level features. These features are used for regression tasks.

The above process is repeated for 90 drug graphs and the feature extraction varies based on the architecture of GNN variants which will be discussed in the model building section. The GNN architecture uses separate adjacency matrices, global features, node features, and edge features for ninety drugs and is further used for building IC50 regression models.

2.2. Building IC50 regression model

The IC50 regression model uses three architectures namely GAT, MPNN, and GIN. These architectures integrate drug SMILES with gene expression thus enhancing and aiding in precise IC50 values prediction. The gene expression is given as global features for building IC50 prediction models. The three models employ distinct architectures to capture different graph perspectives, including feature projection with attention mechanism, user-defined aggregation functions, and permutation invariance. This enhances accuracy and offers diverse insights into molecular graphs.

2.2.1. Graph attention network architecture

The attention mechanism prioritizes relevant components over less relevant ones. As a result, by focusing on the most important parameters, the model can produce more accurate predictions. In the general agreement on trade in services (GATS) scenario, the significance of the connections among nodes in a graph is assessed by utilizing the attention mechanism. Conventional graph convolutional networks (GCNs) employ a preset connection weighting strategy, which might not be the best option for all kinds of graphs. However, depending on the task and network structure, attention mechanisms allow the model to give varying weights to different connections between them. The GAT architecture is given in Figure 3 for a single graph.

In this model, the feature matrix (30 features for each atom) and adjacency matrix are input into the GAT layers. The atom features encoded properties like atom type, formal charge, hybridization, hydrogen bonding, aromatic, degree, number of hydrogens, chirality, and partial charge. GAT layers update each node's features based on neighboring nodes, guided by the adjacency matrix. The main model combines GATLayers, readout, and prediction layers. Each GATLayer applies a graph attention mechanism with fully

connected layers, feature and attention dropouts (0.2), Leaky rectified linear units (LeakyReLU) activation, and residual connections. The first GAT layer transforms 30 features to 64, while the second maintains 64 features. Attention dropout applies to attention weights, and LeakyReLU introduces non-linearity. The adjacency matrix guides node interactions during attention-based updates, with only connected nodes influencing each other. After GAT layers, each atom has an 8-dimensional feature vector used in the readout layer, where a WeightedSumAndMax mechanism computes weights for each atom, emphasizing relevant nodes. The resulting graph-level feature vector that has extracted node features with concatenated gene expression features is input to a multi-layer perceptron (MLP) predictor, starting with a linear layer expanding features from 16 to 128, followed by rectified linear units (ReLU), batch normalization, and a final linear layer producing a single regression output.

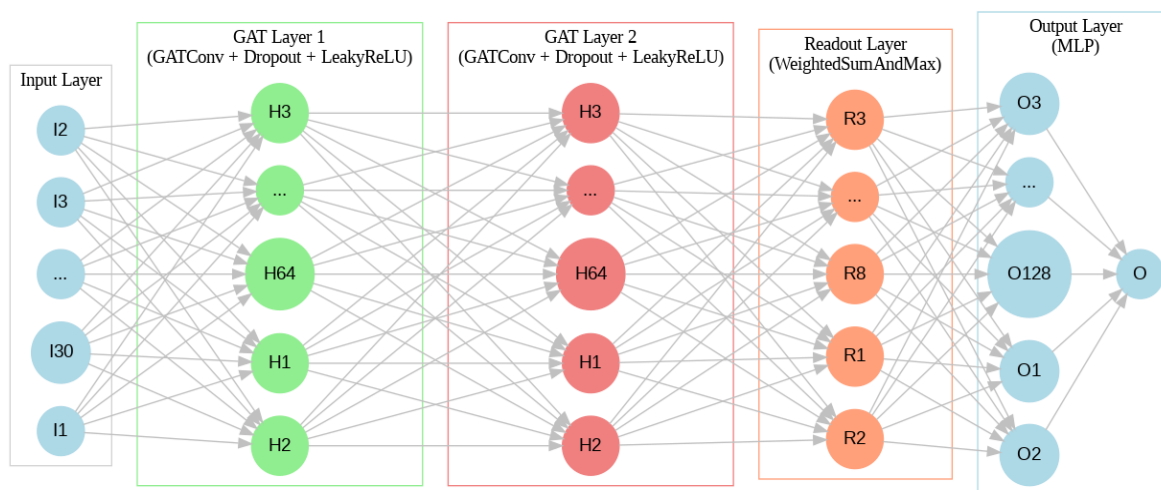


Figure 3. GAT architecture

2.2.2. Message passing neural network architecture

MPNNs are a general framework for GNNs. They use a flexible message-passing mechanism, which can involve passing information between nodes in various ways. User-defined functions and aggregations are utilized to make the message-passing mechanism flexible. GCNs are a specific type of MPNN with a fixed, simple message-passing mechanism. MPNNs are a more general framework that allows for greater flexibility in propagating information across nodes in a graph. The MPNN architecture is given in Figure 4 for a single graph.

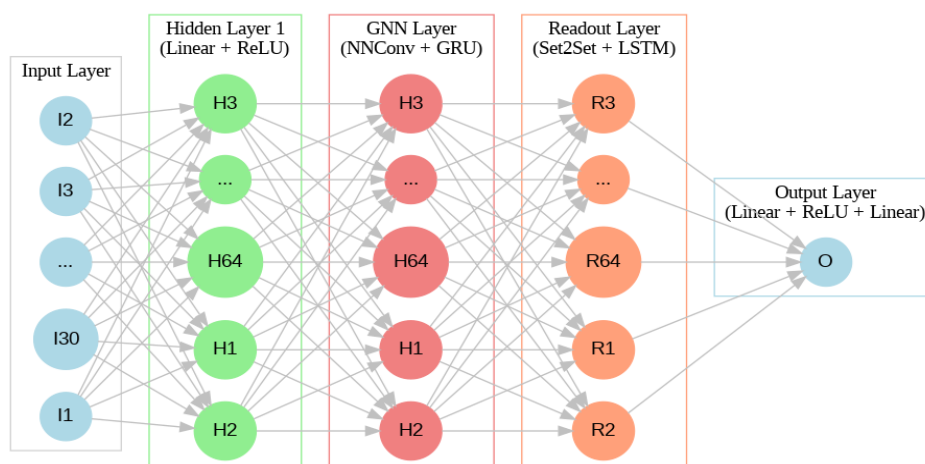


Figure 4. MPNN architecture

The MPNN architecture includes node projection, message passing, a readout layer, and predictor operations. The node projection transforms 30 input node features into 64 features using a linear transformation followed by ReLU activation, projecting them to a higher-dimensional space. The GNN layer employs graph convolution and gated recurrent units (GRU) layers. The graph convolution uses the adjacency matrix to aggregate information from neighboring nodes, where the edge function processes 11 edge features (e.g., bond type, same ring, conjugated, stereo) to modulate message passing. The GRU layer refines 64-node features iteratively, learning dependencies across nodes to capture complex structures. During message passing, node and edge features generate messages, which update node representations through recurrent processing with GRU, maintaining a dynamic state. The Readout layer concatenates gene expression features with node features and aggregates them to form a graph-level representation, using a long short-term memory network (LSTM) (running for six iterations) to refine the global summary. Predictor layers use a dense network with 'Linear' and 'ReLU' transformations to produce the final prediction, such as the predicted drug response.

2.2.3. Graph isomorphism network architecture

GINs are designed to be permutation invariant and can operate on both directed and undirected graphs. They use multiple aggregation steps and apply a shared MLP to aggregate features from the node's neighborhood. Figure 5 provides a simplified representation of the GIN architecture. The GIN combines an input layer with 67 node features, graph isomorphism conv layers, and a readout mechanism.

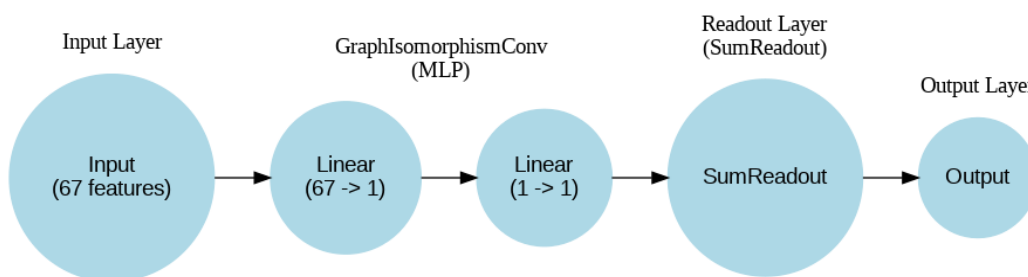


Figure 5. GIN architecture

GraphIsomorphismConv` with layers that involve transformations from `67` features to `1`, and then from `1` to `1`. The `[1]` in 'hidden_dims' indicates that the hidden layer has a size of one, which is consistent with the output of the MLP layers. The GraphIsomorphismConv layer is the key operation that combines the feature matrix and adjacency matrix. The MLP inside this layer takes the node features (from the feature matrix), processes them through linear layers, and then uses the adjacency matrix to aggregate the transformed features from neighboring nodes. The output of the GINConv layer is likely a refined set of node representations that captures information from neighboring nodes in a permutation-invariant manner. The extracted node features are concatenated with gene expression features in the readout layer. After the sum readout is performed that aggregates entire features into a graph-level representation. This aggregation sums up the features of all nodes, which have already been updated considering their neighbors, thus incorporating both the feature matrix and the graph structure defined by the adjacency matrix which is passed to the linear output layer for IC50 prediction.

3. EXPERIMENTS AND RESULTS

DeepChem [33] is a front-end software package specialized in cheminformatics [34] tasks, whereas TensorFlow is the backend that handles model computations. The three distinct IC50 regression models were created using GNN variants in Python through training the adjacency matrix, node features, edge features of drug SMILES, and global features of gene expression. The tests are carried out with different epoch lengths and other hyperparameters, as indicated in Table 2. A single neuron is defined in the output layer for regression purposes. The Adam optimizer is employed here to eliminate errors and improve efficiency.

Table 2. Hyperparameters setting for GNN variants

GAT model		MPNN model		GIN model	
Epoch	500	Epoch	500	Epoch	500
GAT layer	2	MPNN layer	1	GIN layer	1
GAT neuron size	32	MPNN neuron size	32	GIN neuron size	67
Dense layer	1	Dense layer	1	Dense layer	1
Dense layer neurons	64	Dense layer neurons	64	Dense layer neurons	1
Learning rate	0.001	Learning rate	0.001	Learning rate	0.001
Output size	1	Output size	1	Output size	1
Optimizer	Adam	Optimizer	Adam	Optimizer	Adam

The GAT, MPNN, and GIN network is trained iteratively from epoch 10 to 500 at different folds. The GAT, MPNN, and GIN prediction outcomes for MAE, RMSE, and R2 score over numerous epochs are shown in Table 3. The maximum accuracy generated by the GAT is an R2 score of 0.92, with losses of 0.20 for MAE and 0.17 for RMSE. The maximum accuracy generated by the MPNN is an R2 score of 0.85, with losses of 0.26 for MAE and 0.24 for RMSE. The maximum accuracy generated by the GIN is an R2 score of 0.80, with losses of 0.44 for MAE and 0.42 for RMSE. MAE, RMSE, and R2 score are computed for all epochs at intervals of 100. The precision and error rate have continually increased, leading to maximum accuracy in the R2 score and minimal error values for MAE and RMSE. Similarly, the highest accuracy provided by the R2 score is 0.92, indicating a 92% precision rate. The performance findings of the IC50 regression model utilizing drug SMILES and gene expression are provided in Table 3 and illustrated in Figures 6(a) to (c).

Table 3. Performance results of IC50 regression models

		GAT-based IC50 prediction model				MPNN-based IC50 prediction model				GIN-based IC50 prediction model			
7-fold cross validation	Epoch	MAE	RMSE	R2 score	Epoch	MAE	RMSE	R2 score	Epoch	MAE	RMSE	R2 score	
	100	0.48	0.38	0.40	100	0.60	0.49	0.30	100	0.68	0.58	0.25	
	200	0.40	0.30	0.50	200	0.42	0.40	0.41	200	0.52	0.50	0.30	
	300	0.31	0.28	0.56	300	0.43	0.39	0.44	300	0.51	0.48	0.42	
	400	0.32	0.22	0.62	400	0.44	0.33	0.53	400	0.49	0.42	0.52	
	500	0.20	0.17	0.92	500	0.26	0.24	0.85	500	0.44	0.40	0.80	
	Epoch	MAE	RMSE	R2 score	Epoch	MAE	RMSE	R2 score	Epoch	MAE	RMSE	R2 score	
5-fold cross validation	Epoch	MAE	RMSE	R2 score	Epoch	MAE	RMSE	R2 score	Epoch	MAE	RMSE	R2 score	
	100	0.44	0.36	0.30	100	0.57	0.44	0.28	100	0.76	0.60	0.23	
	200	0.40	0.30	0.42	200	0.45	0.41	0.36	200	0.61	0.59	0.37	
	300	0.31	0.28	0.48	300	0.43	0.43	0.41	300	0.47	0.50	0.45	
	400	0.42	0.31	0.60	400	0.40	0.47	0.59	400	0.52	0.49	0.58	
	500	0.32	0.29	0.80	500	0.26	0.35	0.78	500	0.48	0.45	0.75	
	Epoch	MAE	RMSE	R2 score	Epoch	MAE	RMSE	R2 score	Epoch	MAE	RMSE	R2 score	
2-fold cross validation	Epoch	MAE	RMSE	R2 score	Epoch	MAE	RMSE	R2 score	Epoch	MAE	RMSE	R2 score	
	100	0.62	0.67	0.10	100	0.60	0.49	0.20	100	0.88	0.78	0.18	
	200	0.52	0.48	0.20	200	0.56	0.50	0.25	200	0.65	0.61	0.20	
	300	0.49	0.40	0.30	300	0.43	0.39	0.36	300	0.52	0.51	0.34	
	400	0.47	0.39	0.50	400	0.49	0.40	0.54	400	0.57	0.52	0.46	
	500	0.39	0.38	0.70	500	0.34	0.36	0.65	500	0.52	0.49	0.60	

The performance results of the above three GNN variants based IC50 regression models built with drug SMILES and gene expression are compared with each other and also with IC50 regression models built with basic GCN. The highest accuracy is achieved at epoch 500. GAT-IC50 regression model obtains 0.92 R2 score, RMSE 0.17, and MAE 0.20. MPNN-IC50 regression model obtains an R2 score of 0.85, RMSE 0.24, and MAE 0.26. GIN-IC50 regression model obtains an R2 score of 0.80, RMSE 0.40, and MAE 0.44. GCN-IC50 regression model obtains an R2 score of 0.78, RMSE 0.48, and MAE 0.50. The comparative results are also given in Figure 6(d).

The proposed GNN variations that use drug SMILES and gene expression show enhanced IC50 prediction ability. Integrating gene expression data as global traits improves the model's prediction ability. The GNN variant's level of recognition validates its capacity to predict IC50. Compared to other artificial intelligence or deep learning methods, the GNN variants-based IC50 prediction system accurately depicts drug compounds with gene expression. Based on the IC50 model building process GAT performs well, after which MPNN and finally GIN for our dataset. GAT-IC50 regression model uses an additional attention mechanism assigning dynamic weights to node connections. MPNN-IC50 regression model allows flexible message-passing functions to be defined and finally GIN-IC50 regression model achieves permutation invariance in graph data.

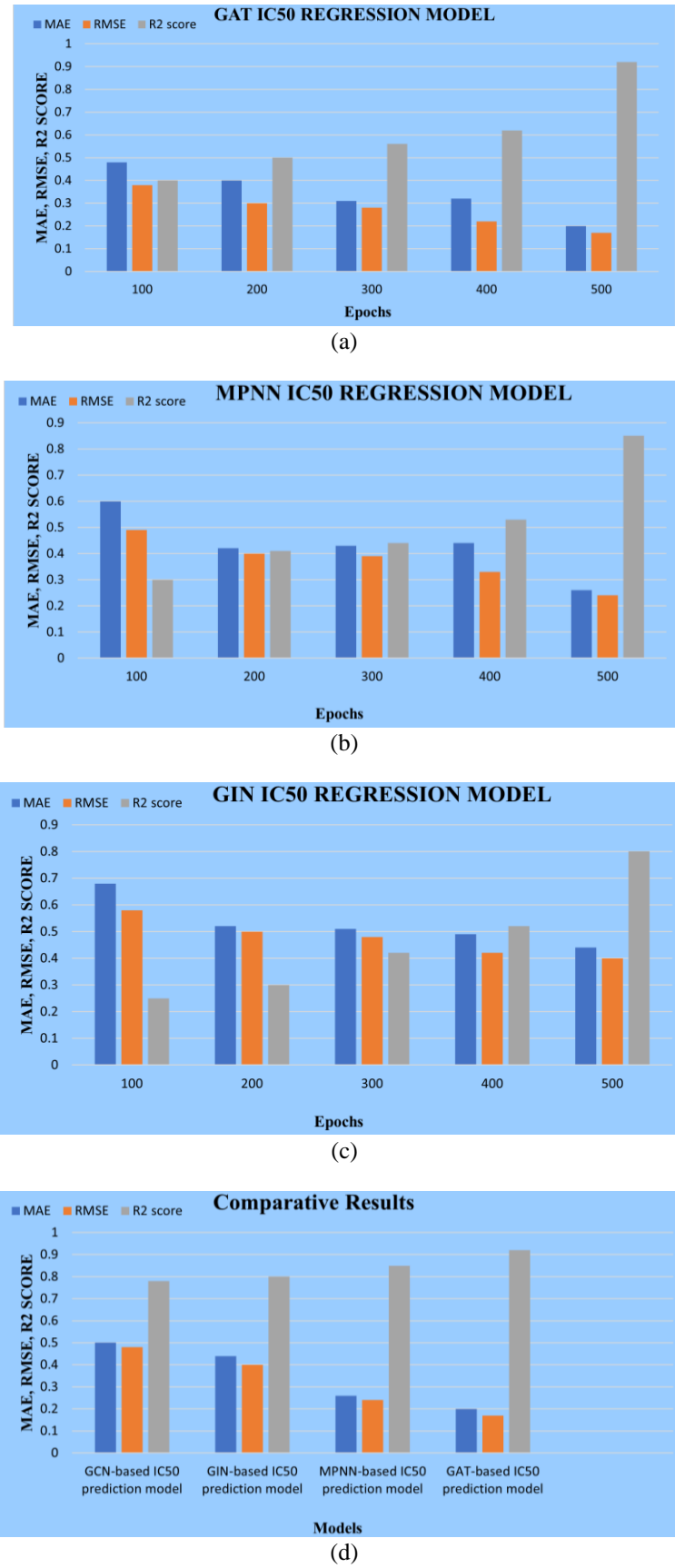


Figure 6. Performance results for; (a) GAT regression model, (b) MPNN regression model, (c) GIN regression model, and (d) comparative results

3.1. Comparative analysis

The comparative analysis of the GNNs IC50 regression model with the previous works like the ML-based IC50 model, GRU-based IC50 regression model [35], and graph convolutional neural network IC50 regression model [32], [35] on evaluation with various datasets for ALS disease is displayed in Table 4. To overcome the issues of huge data size, irrelevant feature selection in ML models, and huge vector size in GRU models, the proposed model uses the advantage of GNN architectures to make better feature extraction on drug SMILES and gene expression of ALS disease. The GAT regression model results in the best accuracy in terms of MAE of 0.20, RMSE of 0.17, and R2 score of 0.92. The MPNN IC50 regression model results in an MAE of 0.26, RMSE of 0.24, and R2 score of 0.85. The GIN IC50 regression model results in an MAE of 0.44, RMSE of 0.40, and R2 score of 0.80. GNNs outperform previous models. Different algorithms were tried like random forest regressor, support vector machine (SVM), multi-layer perceptron regressor (MLPRegressor), GRU, and GCN on drug SMILES data but the proposed method leverages drug SMILES and gene expression data and improves the accuracy by using enhanced GNN algorithms and hyperparameter settings. The proposed method can be used in scenarios where gene expression is considered for disease analysis and drug-gene interactivity is to be learned thereby contributing to drug discovery.

Table 4. Comparative analysis

Datasets	Models used	Results		
		RMSE	MAE	R2 score
Molecular descriptors of 500 drug SMILES	ML model	0.4	0.48	0.58
Drug sequences of 500 drug SMILES	GRU	0.32	0.23	0.63
Drug-induced Gene expressions of 201 drugs	GCNs	0.0038	-	-
Drug graph of 2,100 drugs	GCNs	0.2	0.3	0.73
Drug SMILES and gene expressions of 80 drugs	GCNs	0.18	0.16	0.90

3.2. Discussion

This section addresses the benefits of the proposed GNN variations method over earlier approaches for predicting IC50 in ALS. The limitation of the previous work, the ML-based model was that it was difficult to train the ML model to obtain high robustness and accuracy due to molecular descriptors as training data. Similarly, GRU and graph convolutional neural network models were previously trained with drug SMILES alone. However, analyzing gene expression data necessitates a sophisticated model and a creative methodology. It is accomplished by adding gene expression as a global feature into drug graphs via graph convolutional neural network architecture but with limited generalizability. It is further improved with the proposed enhanced GNN variants GAT, MPNN, and GIN to overcome the above challenges, this research introduces better drug-gene interactivity learning with robustness for 90 drugs. The GNN variants use different feature extraction techniques like attention mechanisms, customized aggregation functions, and permutation invariant representations. The proposed method results are compared with the existing methods like the ML-based IC50 model, GRU-based IC50 regression model [32], and graph convolutional neural network IC50 regression model [32], [33]. Compared to GCN trained using 80 drugs and their gene expression data, the proposed enhanced GNN IC50 regression model achieves higher accuracy and greater generalizability. It attains superior results when compared to the previous works. However, the proposed method does not support datasets like mutation and copy number variation which in the future should approach a different methodology and algorithms for their analysis in drug discovery.

4. CONCLUSION

This study describes the use of innovative GAT, MPNN, and GIN architectures to develop an IC50 regression model using drug SMILES in graph form with gene expression as global characteristics. The model employs GAT, MPNN, and GIN to record structural data and interconnections inside a drug molecule, depicted as graphs. The Uniprot, ChEMBL, and Depmap databases were utilized in this work. The drug SMILES and gene expression data obtained for 90 medications are transformed and employed as an adjacency matrix, node features, and edge features to train the GNN variants. The GNN variants were developed using the DeepChem framework, and the trials were performed with suitable hyperparameter settings. The IC50 prediction model was tested for efficiency using typical metrics and showed promising outcomes in predicting the IC50. In the future, other data modalities can be used for IC50 regression and analysis. This approach enhances understanding of the factors influencing IC50 values and allows for more precise predictions, aiding drug optimization.

ACKNOWLEDGMENTS

We are grateful to PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India, for their tremendous support and resources, which contributed to the success of our research endeavor.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Devipriya Selvaraj	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓
Vijaya M S	✓			✓		✓		✓		✓		✓	✓	
Krishnaveni Sakkarapani	✓			✓		✓		✓		✓		✓	✓	✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.




REFERENCES

- [1] D. He *et al.*, "De novo generation and identification of novel compounds with drug efficacy based on machine learning," *Advanced Science*, vol. 11, no. 11, 2024, doi: 10.1002/adv.202307245.
- [2] P. Masrori and P. Van Damme, "Amyotrophic lateral sclerosis: a clinical review," *European Journal of Neurology*, vol. 27, no. 10, pp. 1918–1929, 2020, doi: 10.1111/ene.14393.
- [3] M. A. van Es *et al.*, "Amyotrophic lateral sclerosis," *The Lancet*, vol. 390, pp. 2084–2098, Nov. 2017, doi: 10.1016/S0140-6736(17)31287-4.
- [4] A. Hekmat, A. A. Saboury, and L. Saso, "Chapter 8.1 - superoxide dismutases inhibitors," in *Metalloenzymes*, Academic Press, 2024, pp. 523–531, doi: 10.1016/B978-0-12-823974-2.00004-8.
- [5] X. Wang, Y. Hu, and R. Xu, "The pathogenic mechanism of TAR DNA-binding protein 43 (TDP-43) in amyotrophic lateral sclerosis," *Neural Regeneration Research*, vol. 19, no. 4, pp. 800–806, 2024, doi: 10.4103/1673-5374.382233.
- [6] Y. Geng and Q. Cai, "Role of C9orf72 hexanucleotide repeat expansions in ALS/FTD pathogenesis," *Frontiers in Molecular Neuroscience*, vol. 17, 2024, doi: 10.3389/fnmol.2024.1322720.
- [7] C. Shum *et al.*, "Mutations in FUS lead to synaptic dysregulation in ALS-iPSC derived neurons," *Stem Cell Reports*, vol. 19, no. 2, pp. 187–195, 2024, doi: 10.1016/j.stemcr.2023.12.007.
- [8] V. Nithianandam, S. Sarkar, and M. B. Feany, "Pathways controlling neurotoxicity and proteostasis in mitochondrial complex I deficiency," *Human Molecular Genetics*, vol. 33, no. 10, pp. 860–871, May 2024, doi: 10.1093/hmg/ddae018.
- [9] H. L. Smith, H. Chaytow, and T. H. Gillingwater, "Excitotoxicity and ALS: New therapy targets an old mechanism," *Cell Reports Medicine*, vol. 5, no. 2, 2024, doi: 10.1016/j.xcrm.2024.101423.
- [10] C. Peggion, T. Calì, and M. Brini, "Mitochondria dysfunction and neuroinflammation in neurodegeneration: who comes first?," *Antioxidants*, vol. 13, no. 2, 2024, doi: 10.3390/antiox13020240.
- [11] W. Wu *et al.*, "Genes in axonal regeneration," *Molecular Neurobiology*, vol. 61, no. 10, pp. 7431–7447, Oct. 2024, doi: 10.1007/s12035-024-04049-z.
- [12] M. Kahl *et al.*, "m6A RNA methylation regulates mitochondrial function," *Human Molecular Genetics*, vol. 33, no. 11, pp. 969–980, 2024, doi: 10.1093/hmg/ddae029.
- [13] C. Hasselgren and T. I. Oprea, "Artificial intelligence for drug discovery: Are we there yet?," *Annual Review of Pharmacology and Toxicology*, vol. 64, pp. 527–550, 2024, doi: 10.1146/annurev-pharmtox-040323-040828.
- [14] S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, and M. J. Ahsan, "Machine Learning in Drug Discovery: A Review," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 1947–1999, Mar. 2022, doi: 10.1007/s10462-021-10058-4.
- [15] K. P. Rahate and R. Mondal, "Applications of AI in drug discovery: its challenges, opportunities, and strategies," *Approaches to Human-Centered AI in Healthcare*, pp. 86–120, 2024, doi: 10.4018/979-8-3693-2238-3.ch005.
- [16] Y. Xia, Y. Wang, Z. Wang, and W. Zhang, "A comprehensive review of molecular optimization in artificial intelligence-based drug discovery," *Quantitative Biology*, vol. 12, no. 1, pp. 15–29, 2024, doi: 10.1002/qub.2.30.




- [17] J. Mao *et al.*, “Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models,” *iScience*, vol. 24, no. 9, 2021, doi: 10.1016/j.isci.2021.103052.
- [18] S. Kausar and A. Falcão, “An automated framework for QSAR model building,” *Journal of Cheminformatics*, vol. 10, 2018, doi: 10.1186/s13321-017-0256-5.
- [19] I. Olier *et al.*, “Meta-QSAR: a large-scale application of meta-learning to drug design and discovery,” *Machine Learning*, vol. 107, no. 1, pp. 285–311, 2017, doi: 10.1007/s10994-017-5685-x.
- [20] Z. Wu *et al.*, “Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets,” *Briefings in Bioinformatics*, vol. 22, no. 4, 2020, doi: 10.1093/bib/bbaa321.
- [21] L. Deng *et al.*, “Artificial intelligence-based application to explore inhibitors of neurodegenerative diseases,” *Front Neurobot*, vol. 14, p. 617327, 2020, doi: 10.3389/fnbot.2020.617327.
- [22] C. Knutson, M. Bontha, J. A. Bilbrey, and N. Kumar, “Decoding the protein–ligand interactions using parallel graph neural networks,” *Scientific Reports*, vol. 12, p. 7624, 2022, doi: 10.1038/s41598-022-10418-2.
- [23] A. Sagingalieva, M. Kordzanganeh, N. Kenbayev, D. Kosichkina, T. Tomashuk, and A. Melnikov, “Hybrid quantum neural network for drug response prediction,” *Cancers (Basel)*, vol. 15, no. 10, 2023, doi: 10.3390/cancers15102705.
- [24] Z. Zuo, P. Wang, X. Chen, L. Tian, H. Ge, and D. Qian, “SWnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures,” *BMC Bioinformatics*, vol. 22, no. 1, pp. 1–16, 2021, doi: 10.1186/s12859-021-04352-9.
- [25] A. Park, Y. Lee, and S. Nam, “A performance evaluation of drug response prediction models for individual drugs,” *Scientific Reports*, vol. 13, 2023, doi: 10.1038/s41598-023-39179-2.
- [26] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph Attention Networks,” in *ICLR 2018 Conference Track 6th International Conference on Learning Representations*, 2017.
- [27] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. V. and G. E. Dahl, “Message passing neural networks,” *Machine Learning Meets Quantum Physics*, 2020, pp. 199–214, doi: 10.1007/978-3-030-40245-7_10.
- [28] S. Wang, X. Su, B. Zhao, P. Hu, T. Bai, and L. Hu, “An improved graph isomorphism network for accurate prediction of drug–drug interactions,” *Mathematics*, vol. 11, no. 18, 2023, doi: 10.3390/math11183990.
- [29] The UniProt Consortium, “UniProt: the Universal Protein Knowledgebase in 2025,” *Nucleic Acids Research*, vol. 53, no. D1, pp. D609–D617, Jan. 2025, doi: 10.1093/nar/gkae1010.
- [30] B. Zdrazil *et al.*, “The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods,” *Nucleic Acids Research*, vol. 52, no. D1, pp. D1180–D1192, Jan. 2024, doi: 10.1093/nar/gkad1004.
- [31] R. Arafeh, T. Shibue, J. M. Dempster, W. C. Hahn, and F. Vazquez, “The present and future of the Cancer Dependency Map,” *Nature Reviews Cancer*, vol. 25, no. 1, pp. 59–73, Jan. 2025, doi: 10.1038/s41568-024-00763-x.
- [32] S. Devipriya and M. S. Vijaya, “Graph convolutional neural network for ic50 prediction model with drug smiles graphs and gene expressions of amyotrophic lateral sclerosis,” *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 1, pp. 133–143, Jan. 2024.
- [33] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing, and Z. Wu, *Deep Learning for the Life Sciences*, O'Reilly Media, 2019.
- [34] M. A. Raslan, S. A. Raslan, E. M. Shehata, A. S. Mahmoud, and N. A. Sabri, “Advances in the Applications of Bioinformatics and Chemoinformatics,” *Pharmaceuticals*, vol. 16, no. 7, p. 1050, Jul. 2023, doi: 10.3390/ph16071050.
- [35] S. Devipriya and M. S. Vijaya, “Graph convolutional neural network for IC50 prediction model using amyotrophic lateral sclerosis targets,” in *International Conference on Data Science and Applications*, 2024, pp. 77–91, doi: 10.1007/978-981-99-7820-5_7.

BIOGRAPHIES OF AUTHORS






Devipriya Selvaraj    received her BCA degree from Dr. G R Damodaran College of Science in 2017. She obtained her M.Sc. degree in Information Technology from PSGR Krishnammal College for Women in 2020. She is pursuing a Ph.D. in Computer Science at PSGR Krishnammal College for Women. Her research interests include artificial intelligence and regression tasks. She can be contacted at email: devipriya041996@gmail.com.



Vijaya M S    is a retired Associate Professor at the Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India, where she has been a faculty member for more than 25 years and completed her Ph.D. in Computer Science from Amrita University, Kerala. Her research interests are primarily in data mining and bioinformatics where she is the author/co-author of over 100 research publications. She can be contacted at email: msvijaya@psgrkcw.ac.in.



Krishnaveni Sakkarapani    completed MCA., M.Phil., and Ph.D., in Computer Science and is currently working as an Assistant Professor in the Data Analytics (PG) Department of PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India. Fourteen years of experience in teaching and published 80+ papers in International Journals and chapters. Also presented 30+ papers at various National and International Conferences. Interested research areas are data mining and warehousing, software engineering, bioinformatics, computer networks, and neural networks. She is a reviewer of National and International Journals. She can be contacted at email: krishnavenis@psgrkcw.ac.in.